IT & C

ISSN 2821 - 8469, ISSN – L 2821 - 8469, Volumul 2, Numărul 1, Martie 2023

Învățarea regulilor de asociere în mineritul datelor

Nicolae Sfetcu

Sfetcu, Nicolae (2023), Învățarea regulilor de asociere în mineritul datelor, *IT & C*, **2**:1, 48-55, DOI: 10.58679/IT87362, https://www.internetmobile.ro/invatarea-regulilor-de-asociere-in-mineritul-datelor/

Publicat online: 02.02.2023

Învățarea regulilor de asociere în mineritul datelor

Nicolae Sfetcu
nicolae@sfetcu.com

Association rule learning in data mining

Abstract

Association rule learning is a rule-based machine learning method for discovering interesting relationships between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interest. Based on the concept of strong rules, association rules were introduced to discover regularities between products in large-scale transaction data recorded by supermarket point-of-sale systems. Such information may be used as a basis for decisions regarding marketing activities, such as, for example, promotional pricing or product placements. In addition to the above example from market basket analysis, association rules are used in many fields today, including web mining, intrusion detection, continuous manufacturing, and bioinformatics.

Keywords: association rules, data mining

Rezumat

Învățarea regulilor de asociere este o metodă de învățare automată bazată pe reguli pentru a descoperi relații interesante între variabilele din bazele de date mari. Este destinată să identifice reguli puternice descoperite în bazele de date folosind unele măsuri de interes. Pe baza conceptului de reguli puternice, s-au introdus reguli de asociere pentru descoperirea regularităților dintre produse în datele tranzacțiilor la scară largă înregistrate de sistemele de puncte de vânzare din supermarketuri. Astfel de informații pot fi folosite ca bază pentru deciziile cu privire la activitățile de marketing, cum ar fi, de exemplu, prețurile promoționale sau plasările de produse. În plus față de exemplul de mai sus din analiza coșului de piață, regulile de asociere sunt folosite astăzi în

multe domenii, inclusiv mineritul web, detectarea intruziunilor, producția continuă și bioinformatica.

Sursa articolului: Drew Bentley, *Business Intelligence and Analytics*. © 2017 Library Press, Licență CC BY-SA 4.0. Traducere și adaptare Nicolae Sfetcu

Cuvinte cheie: reguli de asociere. Mineritul datelor, data mining

IT & C, Volumul 2, Numărul 1, Martie 2023, pp. 48-55
ISSN 2821 - 8469, ISSN – L 2821 – 8469, DOI: 10.58679/IT87362
URL: https://www.internetmobile.ro/invatarea-regulilor-de-asociere-in-mineritul-datelor/

Învățarea regulilor de asociere este o metodă de învățare automată bazată pe reguli pentru a descoperi relații interesante între variabilele din bazele de date mari. Este destinată să identifice reguli puternice descoperite în bazele de date folosind unele măsuri de interes. Pe baza conceptului de reguli puternice, Rakesh Agrawal et al. a introdus reguli de asociere pentru descoperirea regularităților dintre produse în datele tranzacțiilor la scară largă înregistrate de sistemele de puncte de vânzare (POS) din supermarketuri. De exemplu, regula `{ceapa, cartofi}` ⇒ `{burger}` din datele de vânzări ale unui supermarket ar indica faptul că, dacă un client cumpără ceapă și cartofi împreună, este probabil să cumpere și carne de hamburger. Astfel de informații pot fi folosite ca bază pentru deciziile cu privire la activitățile de marketing, cum ar fi, de exemplu, prețurile promoționale sau plasările de produse. În plus față de exemplul de mai sus din analiza coșului de piață, regulile de asociere sunt folosite astăzi în multe domeni de aplicații, inclusiv mineritul utilizării web, detectarea intruziunilor, producția continuă și bioinformatica. Spre deosebire de mineritul secvenței, învățarea regulilor de asociere nu ia în considerare, de obicei, ordinea elementelor fie într-o tranzacție, fie între tranzacții.

Definiție

Exemplu de bază de date cu 5 tranzacții și 5 articole					
ID tranzacție	lapte	pâine	unt	bere	scutece
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Urmând definiția originală a lui Agrawal și colab., problema minării regulilor de asociere este definită astfel:

Fie $I = \{i_1, i_2, ..., i_n\}$un set de atribute *n* binare numite *itemuri*.

Fie $D = \{t_1, t_2, ..., t_m\}$ un set de tranzacții numit *bază de date*.

Fiecare *tranzacție* din *D* are un ID de tranzacție unic și conține un subset de articole din *I*.

O regulă este definită ca o implicație a formei:

$$X \Rightarrow Y$$

unde $X, Y \subseteq I$ și $X \cap Y = \emptyset$.

Fiecare regulă este compusă din două seturi diferite de elemente, cunoscute și sub denumirea de *seturi de itemuri*, X și Y, unde X este numit *antecedent* sau partea stângă (LHS) și Y *consecvent* sau partea dreaptă (RHS).

Pentru a ilustra conceptele, folosim un mic exemplu din domeniul supermarketurilor. Setul de articole este I = {lapte,pâine,unt,bere,scutece} iar în tabel este prezentată o mică bază de date care conține articolele, unde, în fiecare intrare, valoarea 1 înseamnă prezența articolului în tranzacția corespunzătoare, iar valoarea 0 reprezintă absența unui articol în tranzacția respectivă.

Un exemplu de regulă pentru supermarket ar putea fi {unt, pâine} ⇒ {lapte}, ceea ce înseamnă că, dacă se cumpără unt și pâine, clienții cumpără și lapte.

Notă: acest exemplu este extrem de mic. În aplicațiile practice, o regulă are nevoie de un suport de câteva sute de tranzacții înainte de a putea fi considerată semnificativă din punct de vedere statistic, iar seturile de date conțin adesea mii sau milioane de tranzacții.

Concepte utile

Pentru a selecta reguli interesante din setul tuturor regulilor posibile, se folosesc constrângeri asupra diferitelor măsuri de semnificație și interes. Cele mai cunoscute constrângeri sunt pragurile minime de sprijin și încredere.

Fie X un set de articole, X ⇒ Y o regulă de asociere și T un set de tranzacții ale unei baze de date date.

Suport

Suportul este o indicație a frecvenței cu care setul de articole apare în baza de date.

Valoarea suport a lui X în raport cu T este definită ca proporția de tranzacții din baza de date care conține setul de articole X . În formula: supp(X)/N

În exemplul de bază de date, setul de articole {bere,scutece} are suport, deoarece apare în 20% din toate tranzacțiile (1 din 5 tranzacții). Argumentul `supp()` este un set de precondiții și, prin urmare, devine mai restrictiv pe măsură ce crește (în loc să fie mai incluziv).

Încredere

Încrederea este un indiciu al cât de des s-a constatat că regula este adevărată.

Valoarea de încredere a unei reguli, X ⇒ Y, în raport cu un set de tranzacții T, este proporția tranzacțiilor care conține X care conține și Y.

Încrederea este definită ca:

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X).$$

De exemplu, regula {unt,pâine} ⇒ {lapte} are un nivel de încredere de 0,2 > în baza de date, ceea ce înseamnă că pentru 100% dintre tranzacțiile care conțin unt și pâine regula este corectă (100% din cazurile în care un client cumpără unt si paine, cumpără și lapte).

Rețineți că supp(X ∪ Y) înseamnă suportul uniunii elementelor din X și Y. Acest lucru este oarecum confuz, deoarece în mod normal gândim în termeni de probabilități de evenimente și nu de seturi de elemente. Putem rescrie supp(X ∪ Y) ca probabilitate comună $P(E_X \cap E_Y)$, unde E_X și E_Y sunt evenimentele pentru care o tranzacție conține setul de articole X sau, respectiv, Y.

Astfel, încrederea poate fi interpretată ca o estimare a probabilității condiționate $P(E_Y \mid E_X)$, probabilitatea de a găsi RHS a regulii în tranzacții cu condiția ca aceste tranzacții să conțină și LHS.

Creștere

Creșterea (*lift*) unei reguli este definită astfel:

$$\text{lift}(X \Rightarrow Y) = \text{supp}(X \cup Y) / \text{supp}(X) \times \text{supp}(Y)$$

sau raportul dintre suportul observat și cel așteptat dacă X și Y ar fi independenți.

De exemplu, regula {lapte,pâine} ⇒ {unt} are o creștere de $0,2 / 0,4 \times 0,4 = 1,25$.

Dacă regula ar avea o creștere de 1, ar implica faptul că probabilitatea de apariție a antecedentului și cea a consecinței sunt independente una de cealaltă. Când două evenimente sunt independente unul de celălalt, nu poate fi luată în considerare nicio regulă care să implice aceste două evenimente.

Dacă creșterea este > 1, asta ne permite să cunoaștem gradul în care aceste două apariții sunt dependente una de cealaltă și face ca acele reguli să fie potențial utile pentru prezicerea consecințelor în seturile de date viitoare.

Valoarea creșterii este că ia în considerare atât încrederea regulii, cât și setul de date general.

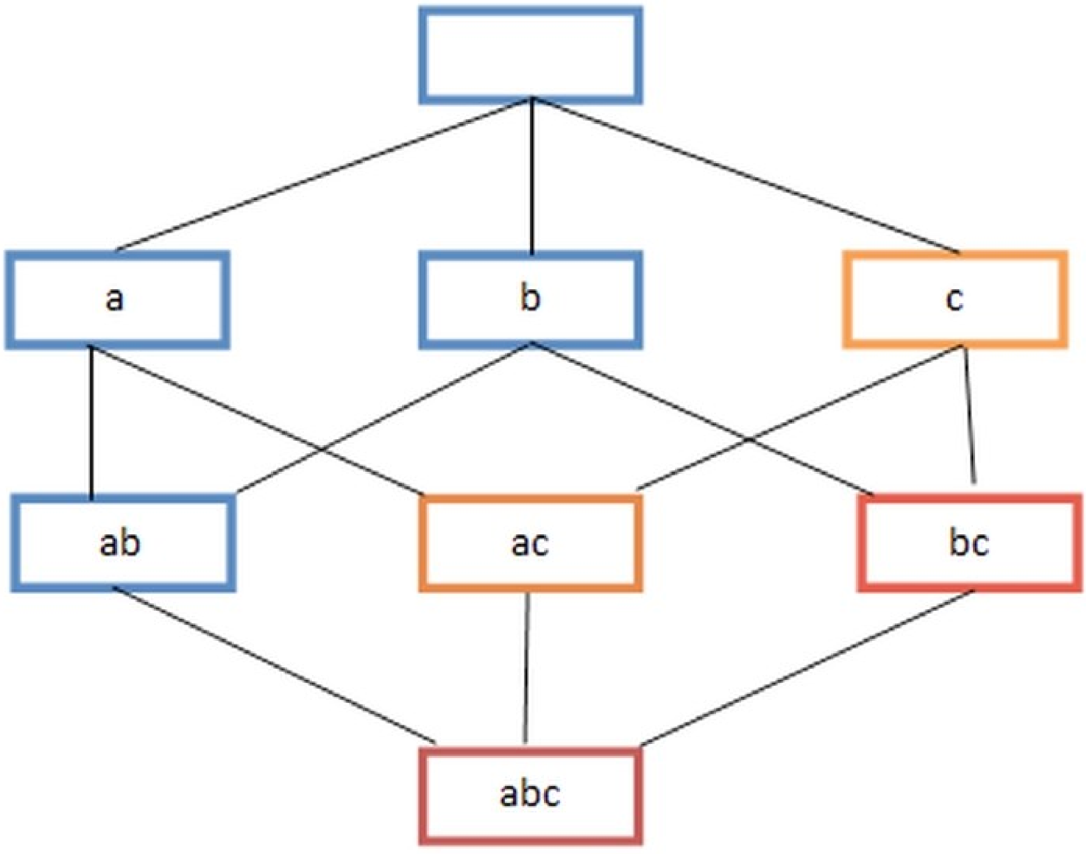
Convingere

Convingerea unei reguli este definită ca $\text{conv}(X \Rightarrow Y) = (1 - \text{supp}(Y))/(1 - \text{conf}(X \Rightarrow Y))$.

De exemplu, regula {lapte, pâine} ⇒ {unt} are o convingere de $(1 - 0,4)/(1 - 0,5) = 1,2$, și poate fi interpretată ca raportul dintre frecvența așteptată pe care X apare fără Y (adică frecvența la care regula face o predicție incorectă) dacă X și Y au fost independenți, împărțit la frecvența observată a predicțiilor incorecte. În acest exemplu, valoarea convingerii de 1,2 arată că

regula {lapte, pâine} ⇒ {unt} ar fi incorectă cu 20% mai des (de 1,2 ori mai des) dacă asocierea între X și Y ar fi pur întâmplătoare.

Proces



(Setul de articole frecvente, unde culoarea casetei indică câte tranzacții conțin combinația de articole. Rețineți că nivelurile inferioare ale rețelei pot conține cel mult numărul minim de articole ale părinților lor; de exemplu. {ac} poate avea numai cel mult min(a, c) elemente. Aceasta se numește **proprietatea de închidere în jos**.)

Regulile de asociere sunt de obicei necesare pentru a satisface un suport minim specificat de utilizator și o încredere minimă specificată de utilizator în același timp. Generarea regulilor de asociere este de obicei împărțită în două etape separate:

1. Se aplică un prag minim de asistență pentru a găsi toate *seturile de articole frecvente* dintr-o bază de date.
2. constrângere minimă de încredere se aplică acestor seturi frecvente de articole pentru a forma reguli.

În timp ce al doilea pas este simplu, primul pas necesită mai multă atenție.

Găsirea tuturor seturilor de articole frecvente într-o bază de date este dificilă, deoarece presupune căutarea tuturor seturilor de articole posibile (combinații de articole). Setul de seturi de

articole posibile este setul de putere peste *I* și are dimensiunea $2^n -1$ (excluzând setul gol care nu este un set de articole valid). Deși dimensiunea setului de putere crește exponențial în numărul de elemente *n* în *I*, este posibilă căutarea eficientă folosind *proprietatea de închidere în jos* a suportului (numită și *anti-monotonitate*) care garantează că pentru un set de articole frecvente, toate subseturile sale sunt de asemenea, frecvente și, prin urmare, pentru un set de articole rar, toate super-seturile sale trebuie să fie, de asemenea, rare. Exploatând această proprietate, algoritmii eficienți (de exemplu, Apriori și Eclat) pot găsi toate seturile de articole frecvente.

Istorie

Conceptul de reguli de asociere a fost popularizat în special datorită articolului din 1993 al lui Agrawal și colab., care a obținut peste 18.000 de citări conform lui Google Scholar, în august 2015, și este, prin urmare, una dintre cele mai citate lucrări din domeniul mineritului de date. Cu toate acestea, este posibil ca ceea ce se numește acum „reguli de asociere” să fie similar cu ceea ce apare în lucrarea din 1966 despre GUHA, o metodă generală de extragere a datelor dezvoltată de Petr Hajek și colab.

O utilizare timpurie (circa 1989) a suportului minim și a încrederii pentru a găsi toate regulile de asociere este cadrul de modelare bazată pe caracteristici, care a găsit toate regulile cu supp(X) și conf(X ⇒ Y) mai mari decât constrângerile definite de utilizator.

Măsuri alternative de interes

Pe lângă încredere, au fost propuse și alte măsuri *de interes* pentru reguli. Câteva măsuri populare sunt:

- Toată încrederea
- Puterea colectivă
- Convingerea
- Pârghia
- Creșterea (numit inițial dobândă)

Mai multe măsuri sunt prezentate și comparate de Tan și colab. și de Hahsler. Căutarea tehnicilor care să modeleze ceea ce a cunoscut utilizatorul (și folosirea acestor modele ca măsuri de interes) este în prezent o tendință de cercetare activă sub numele de „Interesantitate subiectivă”.

Asociații statistice solide

O limitare a abordării standard de a descoperi asocieri este că, prin căutarea unui număr masiv de asocieri posibile pentru a căuta colecții de articole care par a fi asociate, există un risc mare de a găsi multe asocieri false. Acestea sunt colecții de elemente care apar concomitent cu o frecvență neașteptată în date, dar o fac doar întâmplător. De exemplu, să presupunem că luăm în considerare o colecție de 10.000 de articole și căutăm reguli care conțin două articole în partea stângă și 1 articol în partea dreaptă. Există aproximativ 1.000.000.000.000 de astfel de reguli. Dacă aplicăm un test statistic pentru independență cu un nivel de semnificație de 0,05 înseamnă că există doar 5% șanse de a accepta o regulă dacă nu există asociere. Dacă presupunem că nu există asociații, ar trebui să ne așteptăm totuși să găsim 50.000.000.000 de reguli. Descoperirea statistică a asocierilor controlează acest risc, în majoritatea cazurilor reducând riscul de a găsi asocieri false la un nivel de semnificație specificat de utilizator.

Bibliografie

Rakesh Agrawal, Rakesh Agrawal, Tomasz Imielinski, Tomasz Imielinski, Arun Swami, Arun Swami, Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Conference,DOI: 10.1145/170036.170072, în cartea: Proceedings of the 1993 ACM SIGMOD International Conference on Management of DataPublisher: ACM PressEditors: Peter Buneman, Sushil Jajodia

Hâjek, Petr; Feglar, Tomas; Rauch, Jan; and Coufal, David; The GUHA method, data preprocessing and mining, Database Support for Data Mining Applications, Springer, 2004, ISBN 978-3-540-22479-2